Якурнов А.С., аспирант
Трофимов С.П., доц., канд. физ.-мат. наук

СИСТЕМА ПОИСКА ИНФОРМАЦИИ НА ИНТЕРНЕТ-САЙТАХ

В качестве серьезных проблем, которые необходимо решить для реализации системы поиска информации, можно выделить следующие: отсутствие открытых эффективных моделей поиска, работающих в небольших информационных базах; невозможность или сложность настройки систем поиска под требования пользователя.

В данной работе решается задача построения работающей, достаточно эффективной модели поиска информации и предлагается механизм настройки поисковой системы под требования пользователя, или эксперта в определенной области знаний, к результатам поиска.

Моделирование и реальная работа поисковой системы (*searchers.ru*), разработанной А.С. Якурновым, показали, что при серьезном разбросе индексов цитирования (ИЦ) сайтов и из-за различной природы параметров, по которым производится поиск, лучший результат дает мультипликативная нормирующая модель поиска, работающая по принципу улучшения.

Введем вектор параметров $P=(p_1,p_2,p_3,p_4,p_5,p_6,p_7)$, где p_i $i=1...7$ - некоторые параметры просматриваемой страницы: p_1 - число совпадений слов запроса с содержанием страницы; p_2 - число совпадений слов запроса с названием страницы, p_3 - число совпадений слов запроса с ключевыми словами страницы; p_4 - число совпадений слов запроса с описанием страницы; p_6 - индекс цитирования ресурса, которому принадлежит страница; p_7 - количество ссылок на данную страницу, совпадающих с запросом).

Определим релевантность документа $W: P \to R^1$, как отображения вектора параметров P в скаляр.

Получен конечный вид отображения:

$$W(p_1,p_2,p_3,p_4,p_5,p_6,p_7) =$$

$$= \frac{p_1}{N}(c_1+(1-c_1)\sqrt[n_1]{\frac{p_2}{N}})(c_2+(1-c_2)\sqrt[n_2]{\frac{p_3^{R}+p_4^{R}}{2N^{R}}})(1+\lg(\frac{p_6+1}{MAX\_IC}))/(\lg(MAX\_IC))^{K}). \quad (1)$$

Таким образом, модель поиска, то есть вид функции $W(P)$, зависит от множества параметров $\vec{C}=(c_1,c_2,R,n_1,n_2,K)$.

При настройке системы поиска эксперт определяет собственные оценки $W\text{польз}_j$, $j=1,\ldots,m$ для первых m документов, предложенных поисковой машиной. Система определяет свои оценки $W\text{оцен}_j(\vec{C})$ из формулы (1).

В идеале необходимо определить такие параметры $\vec{C}$, при которых

$$\begin{cases} W_1 оцен = W_1 польз \\ W_2 оцен = W_2 польз \\ \ldots \\ W_m оцен = W_m польз \end{cases}.$$

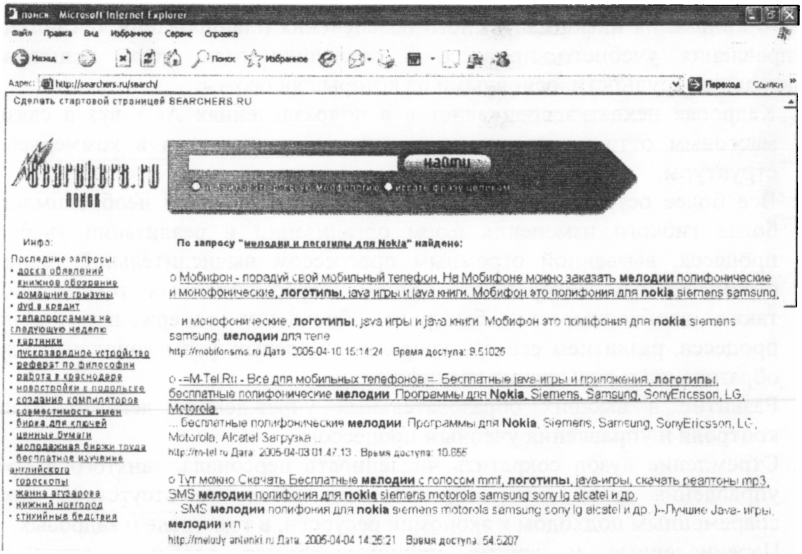
Однако в общем случае, при $m > n$ эта система не разрешается относительно $\vec{C}$, поэтому речь идет о приближении вектора $\vec{W}оцен(\vec{C})$ к $\vec{W}польз$.

Для настройки системы поиска к требованиям эксперта предлагается найти параметры системы $\vec{C}$, при которых отклонение $\left|\vec{W}оцен(\vec{C}) - \vec{W}польз\right|$ минимально.

Таким образом, необходимо решить задачу минимизации целевой функции $F(\vec{C}) = \left|\vec{W}оцен(\vec{C}) - \vec{W}польз\right|$.

После решения этой задачи, например методом сопряженных градиентов, происходит настройка поисковой системы.

Экспериментирование с настройкой системы поиска экспертом показало, что применяемый алгоритм позволяет объективно улучшать результаты поиска в соответствии с требованиями пользователя. Сама поисковая система *searchers.ru* демонстрирует высокое качество поиска информации, хотя предварительно проиндексировано 70.000 стартовых страниц Интернет-сайтов.



Внешний вид системы поиска (сайт *searchers.ru*)